PCT/EP2004/051190

Europäisches Patentamt
European Patent Office
Office européen des brevets

07 08 2004

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

03425436.7

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 03425436.7
Demande no:

Anmeldetag:
Date of filing: 01.07.03
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Semeion
Viale di Val Fiorita 88
00144 Roma
ITALIE

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

An algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions, a method for the cognitive analysis of multidimensional information data based on the said algorithm and a program comprising the said algorithm stored on a recordable support

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/ Classification internationale des brevets:

G06F17/10

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL PT RO SE SI SK TR LI

Semeion

An Algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions a method for the cognitive analysis of multidimensional information data based on the said algorithm and a program comprising the said algorithm stored on a recordable support.

The invention relates to an algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions.

The invention relates particularly to the field of artificial intelligence and the aim is to allow a machine able to carry out computational tasks to analyse complex n-dimensional data in order to represent this data in a two or three dimensional space and so to evaluate this data for cognitive tasks, as for example to create a simplified and representable image of the data or to evaluate the existence of relationships between a group of data records which relationships cannot be represented by exact computable or mathematical functions or for computational tasks in order to solve a problem which is not based on exact mathematical functions.

As it is known nature cannot be always represented by functions having an exact solution or by system of equations having a tall a mathematical solution. In the exact sciences a model may be constructed for simplifying the relationships and helping the

mathematical inspection to be carried out in order to achieve a mathematical or graphic representation. Artificial intelligence is not limited to the analysis and inspection of nature only relatively to exact scientific or technical problems or structures but must be also confronted with social problems which are far most difficult to be represented by the mathematical tools or by exact computable functions. Since artificial intelligence is based on computational machines there is the need of instruments which may help this machines to transform information data in such a way as to be simply handled and used by the machine and in such a way as to allow the machine to generate relationship functions which are easier to handle from the mathematical or computational point of view without distorting or leaving information.

Records of a database, may be represented as points in a space, the position of the points being determined by the variables values which describes the records of the database. In principle the representation may also be reversed in the sense that the variables are represented as points in a space, while the position of each variable is defined by the records. This projection brings certain advantages. As a first technical advantage, certain relationship may be discovered which were hidden in the n-dimensional space of the information data being not intelligible either by human beings nor by machines, since the relative position of the records and/or of the variables in the space where the records or the variables are represented by points is a measure of their similarity or difference. A second technical advantage is that the simplifying of the information

data helps in transforming the data in data which may subjected to a computational evaluation at all and to help the machine for carrying out this computational job in a more rapid and simple way. One might not forget that for mathematical or computational problems there might be theoretically a solution which solution cannot be computed in practice.

The algorithm to which the present invention relates has not only relevance for artificial intelligence, but can also help human intelligence in inspecting and analysing the relationships between information data belonging to a n-dimensional space, where n is bigger then 3 by projecting the data onto a two or three dimensional space. This is a representation which can be understood by human intelligence having its senses constructed to sense a three dimensional or two dimensional space. Thus a representation of data in this space can help human intelligence to understand and find out relationships which could be not be recognised in a four or more dimensional space.

Known algorithm for projecting data form a n-dimensional space into a less dimensional space, and particularly onto a three or two dimensional space uses a predetermined characteristic projection function for computing the position of each point in the projection space. An example for such kind of projection algorithm is the so called Principal Component Analysis, briefly PCA which is described in H.Hotelling "analysis of a Complex of Statistical Variables into Principal Components" J. Educ. Psychol., 24:498-520, 1933. This algorithm provides the steps of defining N factors and N new variables which are orthogonal. Using this base

of new variables a reorganisation of the data is carried out by attempting to put as much information as possible in the first factors under the constraint of linearity. The mapping consist in rewriting the observations/variables using the computed factors and in plotting each one on a two dimensional map using as coordinates the computed factors F1/F2, F3/F4 and so on.

This kind of projection algorithm working only on the base of linear projections determines that some information will be lost during the projection. In order to understand this situation consider a normal projection from a three dimensional space onto a two dimensional space. In a linear projections two points having a certain distance along one of the three dimensions might appear very near if the two dimensional projection space is perpendicular to the third dimension along which the two points are spaced apart. In a very simplified manner this situation takes place using PCA algorithm. The result of the known technique is that in the less dimensional space where the information data has been projected the data relationships is distorted in a dramatic way and the distortion can go so far as to cancel or abnormally enhance relationships between data.

The algorithm according to the present invention has the aim of projecting N-dimensional information data onto a less dimensional, particularly onto a two or three dimensional space without distorting in an excessive manner the relationships between data.

The algorithm according to the present invention has the following steps:

Providing a database of N-dimensional data in the form of records having a certain number of variables.

Defining a metric function for calculating a distance between each record of the database.

Calculating a matrix of distances between each record of the database by means of the metric function defined at the previous step

Defining a n-1 dimensional space in which each record is defined by n-1 coordinates.

Calculating the n-1 coordinates of each record in the n-1 dimensional space by means of an evolutionary algorithm;

Defining as the best projection of the records onto the n-1 dimensional space the projection in which the distance matrix of the records in the n-1 dimensional space best fits or has minimum differences with the distance matrix of the records calculated in the n-dimensional space.

As evolutionary algorithms so called genetic algorithms may be used.

Such kind of algorithm provide new solutions basing on a starting parents population of solutions which may be computed according to various ways such as for example casual attempts. The solutions of the parent populations are combined in such a way that follows the basic combination of genes in genetics thus giving new and different solutions which fitness score, for example in this case the error or difference form the distance matrices of the n- and the n-1 - dimensional spaces is evaluated for giving a certain relevance to the solution which will influence the possibility of combination with other solutions of the new generation for generating a further generation.

This kind of computation makes use of an evolutionary algorithm in order to compute the position of the points in the projection space in such a way to minimize the error with respect to the distances of the points in the original space and is always independent on the specific structure of the information data. Thus on the contrary to the PCA algorithm of the state of the art the algorithm according to the present invention is does not use a predetermined characteristic projection function which computes the position of the points in the projection space.

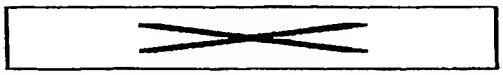
The algorithm according to the present invention combines the projection of the information data with a particular evolutionary algorithm which will be described with greater detail in the following description of the examples.

More in detail the mathematical problem which the present algorithm solves is the following:

Given N points and their distances in a L dimensional space, find into a 2D space the optimal distribution of these points according to the matrix of their constrained distances.

In strict mathematical language the above mentioned problem may be expressed as follows:

Defining a Map Distance in the two dimensional space such as for example:



Where Md is the map distance and i and j are the number of the points and where PX and PY are the coordinates of the point in the two dimensional space.

Defining also a Vector Distance such as

Where Vd is the vector distance I and j are the indices of the different points and V_k are the vector components.

Thus the mathematical problem is to carry out the following optimisation:

Due to the reduction of the number of dimensions in the projection, there might be a situation in which two points might not be separated one from the other if the projection is carried out in a classical way. Thus no exact projection can be carried out form the mathematical point of view if information has to be not distorted or maintained at least partially in the less dimensional space.

The present algorithm solves the above problem by encoding each individual record represented by a point having coordinate X and Y. A set of different X and Y coordinates for each point is defined forming a first population of projections solution onto the less dimensional space, usually a two or three dimensional space.

For each of the projections of this first population the fitness score is calculated by using as the fitness function the matrix of distances of the single points in the originally N dimensional space. The population of projections is then subjected to combination according to the combination rules of the genetic algorithm thus producing a first generation population of projections which comprises X and Y coordinates for the points which are a combination of

the coordinates provided in two projections of the parent generation.

The fitness score of the projections of the first generation is evaluated and again a new generation is formed basing on the first generation.

Using certain combinatory criteria of the projections of the parent generation based on the fitness score of this parent generation the genetic algorithm at each generation the genetic algorithm calculates solution having better fitness scores thus converging against the best solution.

Several genetic or evolutionary algorithm are known see for example, which differ one from the other mostly in the combinatory criteria of the parents in order to generate the next generation of solutions. This criteria relates to the admitted or forbidden "marriages" of two individuals of the parent population and in the mechanism with which the two parents individuals combine their set of data, in this case the different coordinates of the points in the less dimensional map.

As an example a particular genetic algorithm used according to the invention is the so called Genetic Doping Algorithm disclosed in detail in BUSCEMA, 2000: M. Buscema, Genetic Doping Algorithm GenD), Edizioni Semeion, Technical Paper 22e, Rome 2000 and Massimo Buscema & Semeion Group "Reti neurali artificiali e sistemi sociali complessi", Year 199, Edizioni Franco Angeli s.r.l. Milano, Italy, chapter 21, which disclosures are considered to be part of the present specification.

Briefly summarised the GenD algorithm provides for special modified rules for generating the new

individuals of a following generation from the parents population.

As usual in the genetic algorithm, as a first step, GenD calculates the fitness score of each individual, depending on the function that requires optimisation, in this case the distribution function of the data records in the general data set onto the training set and the testing set. The average health score of the entire population is then computed. Average health constitutes the criterion firstly of vulnerability, and secondly of recombination, of all the individuals of the population, for each generation.

All individuals whose health is lower than or equal to the average health of the population are entered in a vulnerability list. This individuals are not eliminated, but continue to take part in the process being only marked out. The number of vulnerable individuals automatically establishes the maximum number of marriages permitted for that generation. The number of possible marriages for each generation thus varies according to the average health of the population. At the third step GenD algorithm couples the individuals. The entire population participate to this possibility. The maximum number of random coupling calls corresponds to half the number of individuals marked out as vulnerable.

For coupling purposes and the generation of children both the candidate individuals must have a fitness value close to the average fitness value of the entire population. Furthermore each couple of individuals may generate off-springs since it is sufficient for marriage that at least one of the two individuals of the couple enjoy health values close to

the health average of the entire population or even higher. According to another recombination rule GenD algorithm does not consider possible marriages between two individuals of which one has a very low health value and the other a very high health value in comparison to the average health value of the population. This means that too weak individuals and too healthy individuals tend not to marry themselves.

Recombination by coupling does not mean classic crossover of the genes of the parents individuals. GenD algorithm effects selective combination of the parents genes by means of two types of recombination. A logic crossover; when repetitions are allowed and an opportunistic crossover; when repetitions are not allowed.

The logic crossover considers four cases:

1. Health of the father and mother are greater than average health of the entire population;
2. Health of both parents is lower than the average health of the entire population;
3. and 4. The Health of one of the parents is less than the average health, while the health of the other of the parents is greater than the average health of the entire population.

If the case 1 does occur than recombination will be effected with a traditional crossover.

If the second case occurs, than the generation of the two children occurs through the rejection of parent's genes.

If case 3 or 4 occur, than the genes of the more healthy parent are transmitted to the children, while the genes of the less healthy parent are rejected.

In the above the definition of rejection does not mean that the rejected genes are cancelled but only that these genes are substituted. Genes substitution is not random but is carried out by means of a sliding window criterion. Each gene may have different genetic options or states. In this case substitution by a sliding window means that the actual rejected gene will be substituted by the same gene but having another state as the original one. So during substitution the criterion used by the GenD algorithm provide only the substitution of the state of that gene which assumes a different state as the gene had in the parent individual.

Relating to the opportunistic crossover, this crossover works when repetition are not allowed. In this case the parents have overlapping genes with respect to a random crossover point. In this case an offspring is generated selecting the more effective gene of the parents. The mechanism is repeated until all the off-springs are completed.

A further criterion of the GenD algorithm rely upon a final opportunity criterion which is a mechanism that enables weak individuals being marked out and having never had the opportunity to be part of a marriage to re-enter the coupling mechanism thanks to a mutation. The number of possible mutations is calculated as the difference of the number of potential marriages and the number of marriages carried out. Mutations occur to those individuals which are present and marked out in the vulnerability list. In this way individuals which had never the opportunity to be part of a generation process is given a final opportunity to enter the evolutionary process.

From the above short explanation of the principal features of this special genetic algorithm it appears clearly that in the GenD algorithm the number of marriages and of mutations are not external parameters, but adaptive self-definable internal variables, taking into account the global tendencies of the population system.

Furthermore It appears also clearly that the basic unit of the GenD algorithm is not the individual, unlike the classic Genetic Algorithm, but the species, which acts on the evolution of individuals in the form of the average health of the entire population of each generation. The feedback loop between individuals and the average health of the population enables the present algorithm to transform in evolution terms the population as a whole from a list of individuals into a dynamic system of individuals.

As a further improvement step of the algorithm according to the present invention, a so called hidden point may be defined. This hidden point whose existence is only guessed is added in the parent population by giving to it position coordinates X_{hi} and Y_{hi} in the projection.

The calculation of the evolutionary algorithm may be carried out in parallel with the hidden point and without the hidden point and the best fit projections obtained by the two parallel calculations may be compared. Hidden points might help in better appreciating the peculiarities of the real positions of the points in the N-dimensional space and so better approximate this positions in the less dimensional projection.

Figure 1 reassumes briefly the mechanism of the present algorithm relatively to the evaluation with and without the hidden unit.

Although the projection carried out with the present algorithm may be on an Euclidean two or three dimensional space, the way the algorithm calculate the projection might be understood as projecting the points for example not in a two dimensional plane but on a two dimensional surface which is somehow curved and nevertheless represented graphically on a plane.

Given a certain database comprising a certain number of records each one characterised by a certain number of variables, the present algorithm might be applied for projecting the database in two different ways.

A first way is to consider the records as being points and the variables as being the coordinates of the points.

The second way is symmetrically reverse this situation by considering the records as variables.

The two spaces are defined as observation and variable spaces and the projections can bring to discovering relations between records and/or between variables.

In the following description of different examples it will be possible to appreciate the effectiveness of the present algorithm and the information which might be recovered by means of the hidden point and also by means of the two projections. From the specific examples it will also appear clearly the technical meaning of the present algorithm which goes beyond the fact of allowing to calculate in a very rapid way the less dimensional map of the points and in a fully

independent way from the structure and meaning of the information data represented by each point or record. In the field of artificial intelligence and thus for example robotics this technical meaning resides in the fact that a computational machine may be able to analyse information data and to recognize or define relationships despite their complexity. The recognitions of relationships of information data in complex problem is important for giving to the machine not only computational power but also allowing the machine to take decisions relating the particular tasks the machine is destined to carry out.

The algorithm according to the present invention can be used for providing method for the cognitive analysis of multidimensional information data.

Provided a database comprising a certain number of records each one representing the relationship between one feature and a certain number of variables, a distance matrix of the records in the N-dimensional space defined by the number of variables characterising each record is calculated according to a certain metric function.

This matrix is taken as the fitness matrix for the projection of the records or the variables into a less dimensional space, particularly a two or three dimensional space using the algorithm described above.

The representation of the records or of the variables in a two or three dimensional space may be used for recognising certain relationships among the records or among the variables.

An example of a particular problem to which the present method can be applied relates to the fact of determining the structure of a certain molecule when

distances between at least some atoms forming the molecules are known. This problem is fo the kind so called not bound optimization problem.

The method can calculate a three or two dimensional projection of the structure of the molecule which may be graphically represented in an intelligible way for human beings. Further more relating to this problem when analysing complex molecules the projection can be carried out by adding the hidden individual which in this case could be an hidden atom and thus offering a tool for inspecting molecular composition and structures of highly complex molecules.

The method according to the present invention being independent on the structure of the information data can handle also not mathematical problems such as for example sociological problems. A first step of giving a numeric scale of evaluation of the different social variables considered might be provided in this case. Normally this is not a critical problem since such kind of variables often are characterised by different state which can be defined by the value true, not true and not present so that the scale in this case could be defined as 1, (-1), and 0.

According to a further feature of the method the algorithm according to the present invention used for projecting the information data from the N-dimensional space onto a less dimensional space, particularly a two or three dimensional space may be applied in combination with other kind of projection algorithm which are somehow more sensitive to the information data structure.

A particular algorithm which can be used in combination with the present projection algorithm is

the so called "SOM" Self Organizing Map algorithm which is a clustering algorithm The SOM is a known algorithm which is described in more details in KOHONEN, 1995: T. Kohonen, Self Organising Maps, Springer Verlag, Berlin, Heidelberg 1995 or Massimo Buscema & Semeion Group "Reti neurali artificiali e sistemi sociali complessi", Year 199, Edizioni Franco Angeli s.r.l. Milano, Italy, chapter 12.

SOM assumes a prior definition of the projection grid and projects codebooks of records in this grid via a competitive algorithm where dominant variables prevail over others.

The SOM projections gives rise to data clusters so called Kohonen units. The SOM algorithm is thus used to perform a first elaboration of the information data, while the algorithm according to the present invention is then used to reproject the Kohonen units emerging from the first elaboration in a coordinated and more detailed manner on its own map.

This procedure allows to take advantage of the peculiarity of the SOM algorithm to consider the significance of the variables and to take also advantage of the features of the present projection algorithm which can evaluate the fitness score of the projection performed referred to the fitness function which is the distance matrix of the points representing the information data I the N-dimensional space and which can also consider hidden units. Thus the reproduction accuracy in the less dimensional space is ensured and a more complex projection which can provide for greater information is performed. Figure 2 schematically illustrates the combination of SOM algorithm with the present algorithm. Thanks to the

present algorithm the advantages of the SOM algorithm are combined with the fact that the present algorithm can dynamically deform the original projections space by hidden units increasing the reconstruction accuracy of the projection.

Figure 1 illustrate a diagram of the architecture of the projection algorithm according to the present invention.

Figure 2 illustrates a diagram of the combination of the algorithm according to the present invention with a SOM algorithm.

Figure 3 is a database of highway distances between Italian cities according to example I.

Figure 4 is the two dimensional projection map of the cities of the database according to the database of figure 3.

Figure 5 is a database of flight distances between US cities according to example II.

Figure 6 is the two dimensional projection map of the cities of the database according to the database of figure 5.

Figure 7 is a database of the European Countries Food Consumption in 1994 according to example III

Figure 8 is the two dimensional diagram of the projection of database of figure 7 in the variable space which has been elaborated by the algorithm according to the present invention.

Figure 9 is the two dimensional diagram of the projection of database of figure 7 in the observation space which has been elaborated by the algorithm according to the present invention.

Figure 10 is a table of 13 variables and their complement of a fourth example.

Figure 11 is the result of the projection of the variables according to the table of figure 10 on a two dimensional plane with the algorithm according to the invention.

Figure 12 illustrates the connection between the variables and the complements on the map according to figure 1.

Figure 13 illustrates a fifth method according to the present invention in which the projection algorithm is used in combination with a so called Self Organising Map.

Figure 14 is a further diagram illustrating the method according to figure 14 and the table of codebooks prototypes of the groups of variables defined by the projection of the database on the two dimensional map.

Example I

In figures 3 and 4 there is described a first example of dataset and two dimensional mapping using the algorithm according to the present invention.

A dataset comprising ten Italian cities and their highway distances is provided. The highway distances are not true two dimensional distances in an Euclidean Space, since every highway distance has three kind of alteration namely: a longitudinal alteration, an altitude alteration and a structural alteration. Thus creating a two dimensional map of the cities where the cities are placed considering only the highway distances using a linear algorithm would determine a distortion of the position of the cities with respect to their real relative position.

The city of Arezzo is not given to the algorithm with its distances from the other cities, but free

determinable distances values are given to the algorithm so that the algorithm is called to look for a hidden city of which the existence is assumed and of witch the position is not known.

A first randomized distance value of the hidden city can be given in the distance matrix for the hidden city so that the algorithm can be initialised and can start to correct the randomized initial position of the hidden city. As it appears clearly from figure 4 a comparison of the map drawn by the algorithm according to the invention with a geographical map allows to identify the hidden city as the city of Arezzo.

Using the algorithm according to the present invention which carries out a non linear projection also the other cities being defined by their distances in the database are placed onto the two dimensional map by optimising their relative position with respect to the matrices of their relative distances. The distortion relative to the real position is very low and the solution is illustrated in figure 4.

EXAMPLE II

Example II is a similar mapping problem as example I. In this case the database comprises twelve US cities and their relative flight distances. No hidden unit has been provided.

Also in this case the flight distances are affected by alterations similar to example I.

Also in this case a linear projection of the cities onto a two dimensional map would not take correctly care of the above mentioned alterations and the position of the cities on the map would be distorted relatively to reality.

The result obtained by the present algorithm is a map which is illustrated in figure 6 and where the positioning of the cities has an error of only 3,07% with respect to the matrix of distances while the positions of the cities is very close to the real geographical position.

EXAMPLE III

Example III is more complex one. The database relates to the European Countries Food Consumption in nineteen ninety four. It comprises nine variables relating to the food kind, namely: cereals, rice, potatoes, sugar , vegetables, meat, milk, butter, eggs.

Sixteen observations where made relating to sixteen countries, namely: Belgium, Denmark, Germany, Greece, Spain, France, Ireland, Italy, Netherlands, Portugal, Great Britain, Austria, Finland, Island, Norway, Sweden.

The database was evaluated with the algorithm according to the present invention and the map according to figure 8 was obtained.

In the two dimensional map the circles indicates geographical areas to which the countries belong. The projection carried out by the present algorithm has shown that there are different groups of countries having similar food consumption and which countries belong to the same geographical area. Furthermore, the two dimensional projection has also highlighted that Ireland has a food consumption behaviour which is very different from that of all the other countries and particularly from the countries of the geographical area to which it belongs.

Figure 9 illustrates the projection of the database made by considering the records as variables,

i.e. the observation countries as variables which has been defined as the observation space. The projection was also carried out by means of the algorithm according to the present invention and the map of figure 9 indicates also a relation which was not apparent from the database.

From the above it appears clearly that the algorithm according to the present invention carries out a projection which due to its non linearity does not lead to hidden information. The PCA algorithm needs to illustrate the information data onto two different maps for not loosing information, while the projection according to the present algorithm does not hide information and relationships between the data.

Returning to the capability of the present mapping algorithm in finding out hidden units this capability can be used for solving further technical problems as for example for drawing two or three dimensional maps of complex molecules also in the case where the list of atoms is incomplete or where the matrix of the distances is incomplete.

It has to be notices that as disclosed above, the database can also be incomplete, this means that despite knowing the presence of certain atoms the distances of this ones may not be known. Thanks to the ability of considering hidden units the algorithm according to the present invention can place the known atom of which the distances where not known in the distance matrix in a correct or most probable position relatively to the other atoms of the molecule.

According to another way of using the capability of considering hidden units relating to this last example, the algorithm according to the present

invention is also capable of considering the presence of unknown atoms in a molecule of which the composition is not completely known and further to this the algorithm is also capable of producing an hypothesis about the most probable position of this atom relatively to the other known atoms thus helping the further study of the molecular structure.

Example IV

In figure 10 the table of 13 variables and their complement is illustrated. The 13 variables relates to anagraphic data and medical data of a certain number of individuals, more precisely of 117 individuals. The aim is to analyse the database in order to find out relations which are somehow connected to the Alzheimer disease or to the probability of developing the Alzheimer disease. Starting from the 13 variables the complement of this variables are defined. The complement being a complementary value of the variables.

Using the aforementioned database the data has been projected onto a two dimensional plane. The result is illustrated in figure 11. From this map the following conclusion can be drawn: The more two variables are nearest, more their information is high and therefore the two variables are similar.

In figure 12 the connection lines between each variable on the map and each complementary variable has been drawn in order to establish their relative distance. From the mathematical point of view it can be demonstrated that the more the connecting segment is

long the more the variable is significative in the database since his standardized variance is bigger.

EXAMPLE V

Figure 13 illustrates a diagram of a combined projection algorithm comprising two different algorithm one of which is a projection algorithm according to the present invention.

A database of different variables for a certain number of individuals comprises 19 variables of medical, anagraphical and social kind. The records of the database are elaborated with an algorithm known as Self Organising Map (SOM). This algorithm clusters the records into cells or units The database is an enlarge version of the one of Example IV.

The algorithm according to the present invention is applied to the units computed by the SOM in order to distribute the said units and the records clustered in it in an optimal way on a two dimensional map. Codebooks prototypes can be computed as the average of the codebooks of each unit taking part to a group.

The groups of units on the two dimensional map created by the projection algorithm are evaluated by means of their clustering on the map. It appears evident that the projection algorithm according to the invention will stress the existence of a fourth group.

Figure 14 illustrates a diagram in which starting form the database comprising the said 19 variables and subjecting the database to the SOM and afterwards to the projection algorithm four groups are generated on the projection map each group having is

specific codebooks prototypes which are listed on the right table.

The variables considered are variables which can be involved in some way with the Alzheimer disease. The number of subjects considered has been of 117 patients. The different groups are characterised by different percentage of patients having developed the Alzheimer disease. The codebooks prototypes can give insight in the relevance of certain medial variables and/or certain anagraphic variables and/or certain social variables for determining the risk of developing the Alzheimer disease by an individuum.

It is interesting to notice that the age is not relevant while social variables such as intellectual level or level of schooling, physical exercise and other variables attaining to the behaviour has a high influence in differentiating the four groups and thus the risk of developing Alzheimer disease. With increasing level of schooling and/o with increasing level of physical exercise and with increasing educational and cultural level the percentage of individual having developed the disease becomes lower, despite the presence of certain pathological variables or medical variables which seems not to be relevant for differentiating the groups one from the other.

From the above mentioned projection different suggestion may be extrapolated:

- Alzheimer disease at histological level starts independently from Tangles in Hippocampus or Plaques in NeoCortex, and arrives to Tangles in NeoCortex passing through Plaques in Hippocampus with different transition probabilities. This suggestion is supported by evidences coming from

the projection algorithm according to the present invention and SOM Systems.

- Severe Braak Stages are related to two different and unrelated pathologies (evidences supported by SOM System).
- Plaques in NeoCortex and Tangles in Hippocampus distribution are connected with two different kind of subjects in SOM System
- MMSE, ADL, BOSTON, and CNPR are strongly connected among them, in the same way that WRCL and VRBF are connected to each other. Evidences supported by the mapping through the projection algorithm according to the present invention which puts these two groups of tests in two different areas.
- Education Years are strongly connected with the Alzheimer disease pathology features (evidences supported by the algorithm according to the present invention).
- The integrated use of different Unsupervised Organisms, allows the identification of four natural clusters of subjects with specific codebooks prototype.

As a device for carrying out the above mentioned method a conventional computer may be used in which the method or the algorithm is loaded as a program. Alternatively the device can be a machine such as a robot or the like having a central computer in which the method and/or the algorithm are loaded as an executable program by the central computer and which program has interfaces with decisional programs relating the functions of the machine or the robot

and/or with learning programs of the machine or of the robot.

CLAIMS

1. An Algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions comprising the following steps:

Providing a database of N-dimensional data in the form of records having a certain number of variables.

Defining a metric function for calculating a distance between each record of the database.

Calculating a matrix of distances between each record of the database by means of the metric function defined at the previous step

Defining a n-1 dimensional space in which each record is defined by n-1 coordinates.

Calculating the n-1 coordinates of each record in the n-1 dimensional space by means of an evolutionary algorithm;

Defining as the best projection of the records onto the n-1 dimensional space the projection in which the distance matrix of the records in the n-1 dimensional space best fits or has minimum differences with the distance matrix of the records calculated in the n-dimensional space.

2. An algorithm according to claim 1 in which a database is provided in which the distances between the records are already contained.

3. An algorithm according to claim 1 or 2, characterised in that as an evolutionary algorithm a so called genetic algorithms is used.

4. An algorithm according to one or more of the preceding claims, characterised by the following steps:

encoding each individual record or variable represented by a point having coordinate X and Y;

defining a set of different X and Y coordinates for each point forming a first population of projections solution onto the less dimensional space, usually a two or three dimensional space.

Calculating the fitness score for each of the projections of this first population by using as the fitness function the matrix of distances of the single points in the originally N dimensional space;

Subjecting the population of projections to combination according to certain combination rules thus producing a first generation population of projections which comprises X and Y coordinates for the points which are a combination of the coordinates provided in two projections of the parent generation;

Calculating the fitness score of the projections of the first generation and forming again a new generation basing on the first generation.

5. An Algorithm according to one or more of the preceding claims, characterised in that the genetic algorithm is the so called GenD algorithm.

6. An algorithm according to one or more of the preceding claims, characterised in that a hidden point can be defined, which corresponds to a hidden record or a to a hidden variable and whose existence is only guessed, the said hidden point is added in the parent population by giving to it position coordinates X_{hi} and Y_{hi} in the projection.

7. An algorithm according to claim 6, characterised in that the calculation of the evolutionary algorithm is carried out in parallel with the hidden point and without the hidden point and the best fit projections obtained by the two parallel calculations is compared.

8. An algorithm according to one or more of the preceding claims characterised by the further steps of providing a database comprising a certain number of records each one characterised by a certain number of variables,

elaborating the database alternatively or in parallel according to two ways:

a first way by which the records are considerd as being points and the variables as being the coordinates of the points.

A second way by which the variables are considered as being points and the records are the coordinates.

9. An algorithm according to one or more of the preceding claims, comprising a further different algorithm treating the database as a pre or post processing phase.

10 An algorithm according to claim 9, characterised in that the database is processed in a preventive stage by means of a Self Organising Map algorithm, the clusters formed by this algorithm in the different units being subjected to a projection by means of the algorithm according to one or more of the proceeding claims.

11. A method for the cognitive analysis of multidimensional information data comprising the following steps:

providing a database with a certain number of records each one comprising a certain number of variables and which are relative to a N-dimensional space;

projecting the database considering the record as points and the variables as coordinates or the

variables as points and the records as coordinates onto a space having a reduced number of dimension relatively to the space N dimensional space;

the projection being carried out by means of an Algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions comprising the following steps:

Calculating a matrix of distances between each point defined by a record or a variable of the database by means of a metric function;

Defining a n-1 dimensional space in which each point represented by a record or a variable is defined by n-1 coordinates.

Calculating the n-1 coordinates of each point in the n-1 dimensional space by means of an evolutionary algorithm;

Defining as the best projection of the points onto the n-1 dimensional space the projection in which the distance matrix of the points in the n-1 dimensional space best fits or has minimum differences with the distance matrix of the points calculated in the n-dimensional space.

12. A method according to claim 11 characterised in that a database is provided in which the distances between the records are already contained.

13. A method according to claims 11 or 12, characterised in that as an evolutionary algorithm a so called genetic algorithms is used.

14. A method according to one or more of the preceding claims 11 to 13, characterised by an algorithm executing the following steps:

encoding each individual record or variable represented by a point having coordinate X and Y;

defining a set of different X and Y coordinates for each point forming a first population of projections solution onto the less dimensional space, usually a two or three dimensional space.

Calculating the fitness score for each of the projections of this first population by using as the fitness function the matrix of distances of the single points in the originally N dimensional space;

Subjecting the population of projections to combination according to certain combination rules thus producing a first generation population of projections which comprises X and Y coordinates for the points which are a combination of the coordinates provided in two projections of the parent generation;

Calculating the fitness score of the projections of the first generation and forming again a new generation basing on the first generation.

15. A method according to one or more of the preceding claims 11 to 14, characterised in that the evolutionary algorithm is the so called GenD algorithm.

16. A method according to one or more of the preceding claims 11 to 15, characterised in that a hidden point represented by a hidden record or a hidden variable can be defined, which corresponds to a hidden point on the map and whose existence is only guessed, the said hidden point being added in the parent population by giving to its position coordinates X_{hi} and Y_{hi} in the projection.

17. A method according to claim 16, characterised in that the calculation of the evolutionary algorithm is carried out in parallel with the hidden point and without the hidden point and the best fit projections obtained by the two parallel calculations is compared.

18. A method according to one or more of the preceding claims characterised by the further steps of providing a database comprising a certain number of records each one characterised by a certain number of variables,

elaborating the database alternatively or in parallel according to two ways:

a first way by which the records are considerd as being points and the variables as being the coordinates of the points.

A second way by which the variables are considered as being points and the records are the coordinates.

19. A method according to one or more of the preceding claims 11 to 18, comprising a further different algorithm treating the database as a pre or post processing phase.

20 A method according to claim 19, characterised in that the database is processed in a preventive stage by means of a Self Organising Map algorithm, the clusters formed by this algorithm in the different units being subjected to a projection by means of the algorithm according to one or more of the proceeding claims.

21. A method according to one or more of the preceding claims 11 to 20, characterised in that the clustering or distance of the points on the map on which the database has been projected is used as a measure of similarity of the records or of the variables related to the said point.

22. A method according to one or more of the preceding claims, characterised in that a database comprising a certain number of records each one being

related to a certain number of variables is provided, to the said database being further added the complementary variables to the variables originally provided and the said integrated database being subjected to projection onto a less dimensional space, particularly on a two or three dimensional space;

The distance in the map between each variable and its complementary variable being used as a measure for the relevance of said variable in the database.

23. A method according to one or more of the preceding claims 11 to 22, characterised in that it is a method for generating two dimensional maps of geographic sites starting form a database comprising relative distances of the sites.

24. A method according to one or more of the preceding claims 11 to 22, characterised in that it is a method for representing the structure of a molecule onto a three dimensional or a two dimensional steps by indicating only the relative distances of the atoms of the molecule.

25. A method according to one or more of the preceding claims 11 to 24, characterised in that it is a method for finding out the presence and/or the position of a unknown or hidden atom in the structure of the molecule.

26 A method according to one or more of the preceding claims, characterised in that it is a method for evaluating the relevance of certain variables in determining a certain pathological status of individuals and a method for defining prototypes of individuals relatively to the variables of the database and their probability of developing a certain disease.

27. A method according to one or more of the preceding claims characterised in that it is a method for analysing the probability of an individual of having or of developing the Alzheimer disease.

28. A method according to one or more of the preceding claims 11 to 27, characterised in that it is in the form of a program saved on a removable support.

29 An Algorithm according to one or more of the preceding claims 1 to 10, characterised in that it is in the form of a program saved on a removable support.

ABSTRACT

An Algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions a method for the cognitive analysis of multidimensional information data based on the said algorithm and a program comprising the said algorithm stored on a recordable support.

An Algorithm for projecting information data belonging to a multidimensional space into a space having less dimensions comprising the following steps: Providing a database of N-dimensional data in the form of records having a certain number of variables; Defining a metric function for calculating a distance between each record of the database; Calculating a matrix of distances between each record of the database by means of the metric function defined at the previous step; Defining a n-1 dimensional space in which each record is defined by n-1 coordinates; Calculating the n-1 coordinates of each record in the n-1 dimensional space by means of an evolutionary algorithm; Defining as the best projection of the records onto the n-1 dimensional space the projection in which the distance matrix of the records in the n-1 dimensional space best fits or has minimum differences with the distance matrix of the records calculated in the n-dimensional space.

The Method and the program apply the aforementioned algorithm.

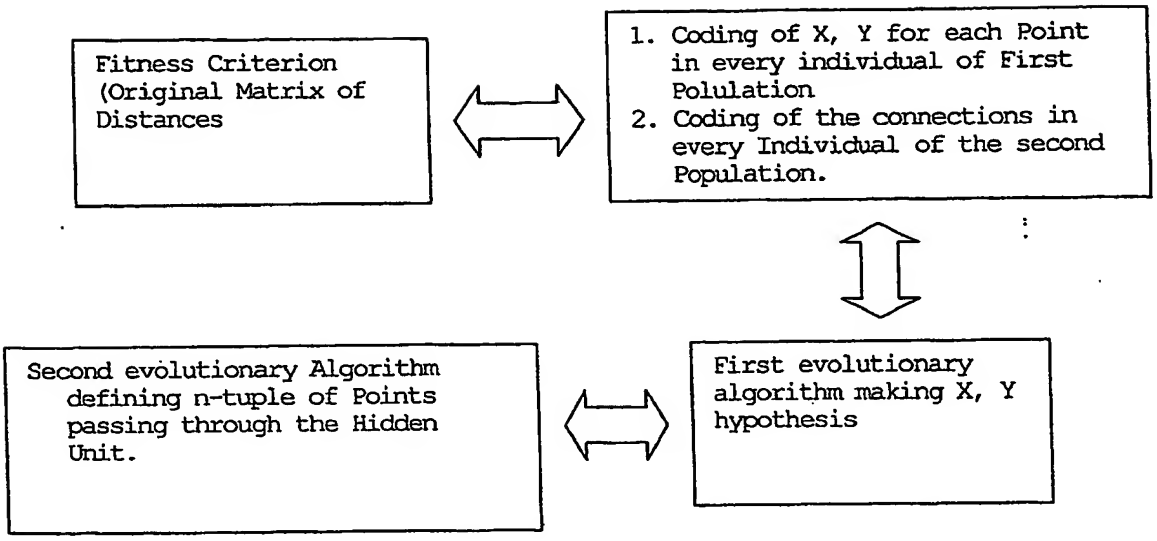

Fitness Criterion (Original Matrix of Distances
1. Coding of X, Y for each Point in every individual of First Polulation
2. Coding of the connections in every Individual of the second Population.
Second evolutionary Algorithm defining n-tuple of Points passing through the Hidden Unit.
First evolutionary algorithm making X, Y hypothesis

Fig. 1

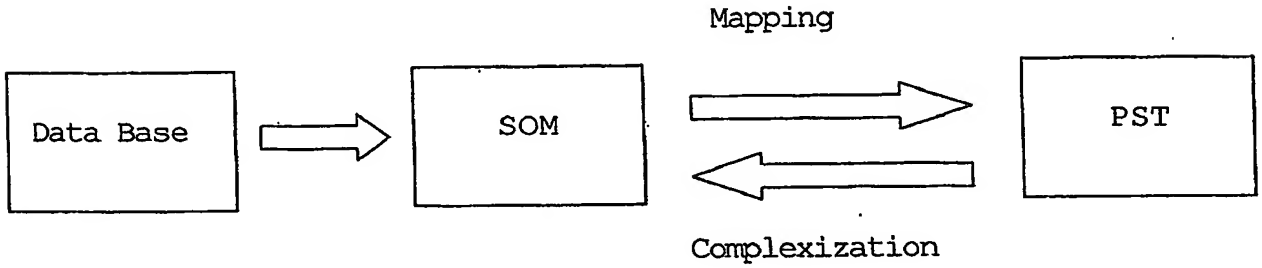

Mapping
Data Base
SOM
PST
Complexization

Fig. 2

Example (1)

HIDDEN	Hidden: Alessandria	Ancona	Aosta	Arezzo	Ascoli	Asti	Avellino	Bari	Belluno	Benevento
Alessandria	0									
Ancona	465	0								
Aosta	165	617	0							
Arezzo	389	192	550	0						
Ascoli	576	122	728	249	0					
Asti	37	491	159	420	602	0				
Avellino	824	437	985	456	365	855	0			
Bari	919	465	1071	661	400	945	208	0		
Belluno	441	454	534	426	565	468	861	908	0	
Benevento	805	395	966	431	323	836	42	197	842	0

Highway Distances in a geographic space between 10 Italian Cities (in Km)
Every highway has three types of alteration in a 2D Euclidean space:

1) A longitudinal alteration
2) An altitude alteration:
3) A structural alteration

Fig. 3

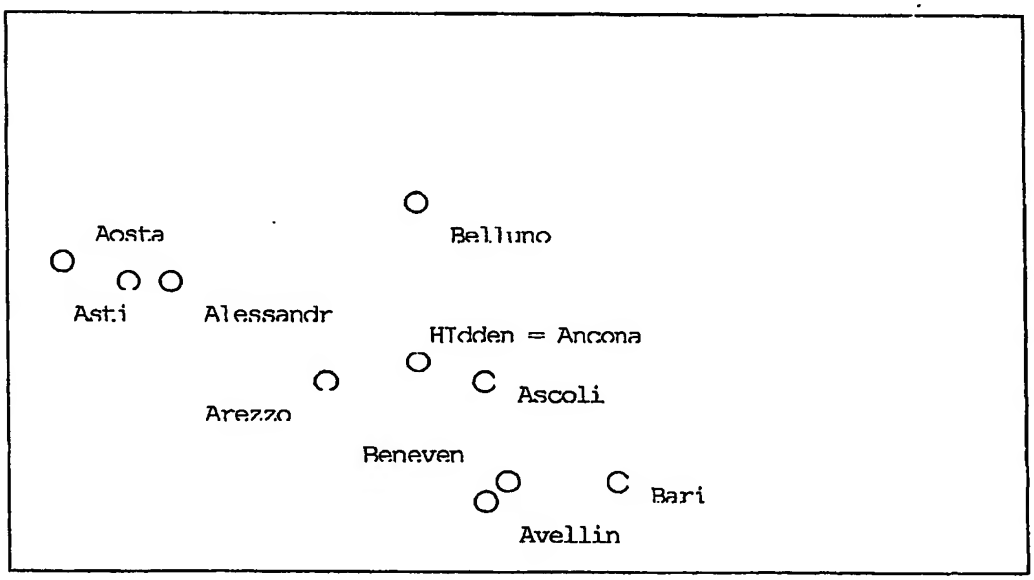


Fig. 4

Example (2)

	LA	NY	BOSTON	DETROIT	BUFFALO	PITTSBURG	CHICAGO	SAINT_LOUIS	CINCINNATI	DALLAS	ATLANTA	MEMPHIS
LA	0											
NY	5600	0										
BOSTON	6109	509	0									
DETROIT	4582	1145	1527	0								
BUFFALO	5091	764	1018	509	0							
PITTSBURG	4836	764	1145	509	382	0						
CHICAGO	4073	1655	2036	509	1018	891	0					
SAINT_LOUIS	3564	2036	2418	1018	1527	1273	636	0				
CINCINNATI	4327	1273	1655	382	764	509	509	764	0			
DALLAS	2800	2927	3436	2036	2545	2291	1655	1018	1782	0		
ATLANTA	4327	1527	2036	1145	1400	1018	1145	1018	764	1527	0	
MEMPHIS	3564	2164	2545	1273	1782	1400	1018	382	891	891	764	0

Flight Distances in a geographic space between 12 USA Cities (in miles)
Every air route has three types of alteration in a 2D Euclidean space:

1) A longitudinal alteration
2) An altitude alteration:
3) A structural alteration

Fig. 5

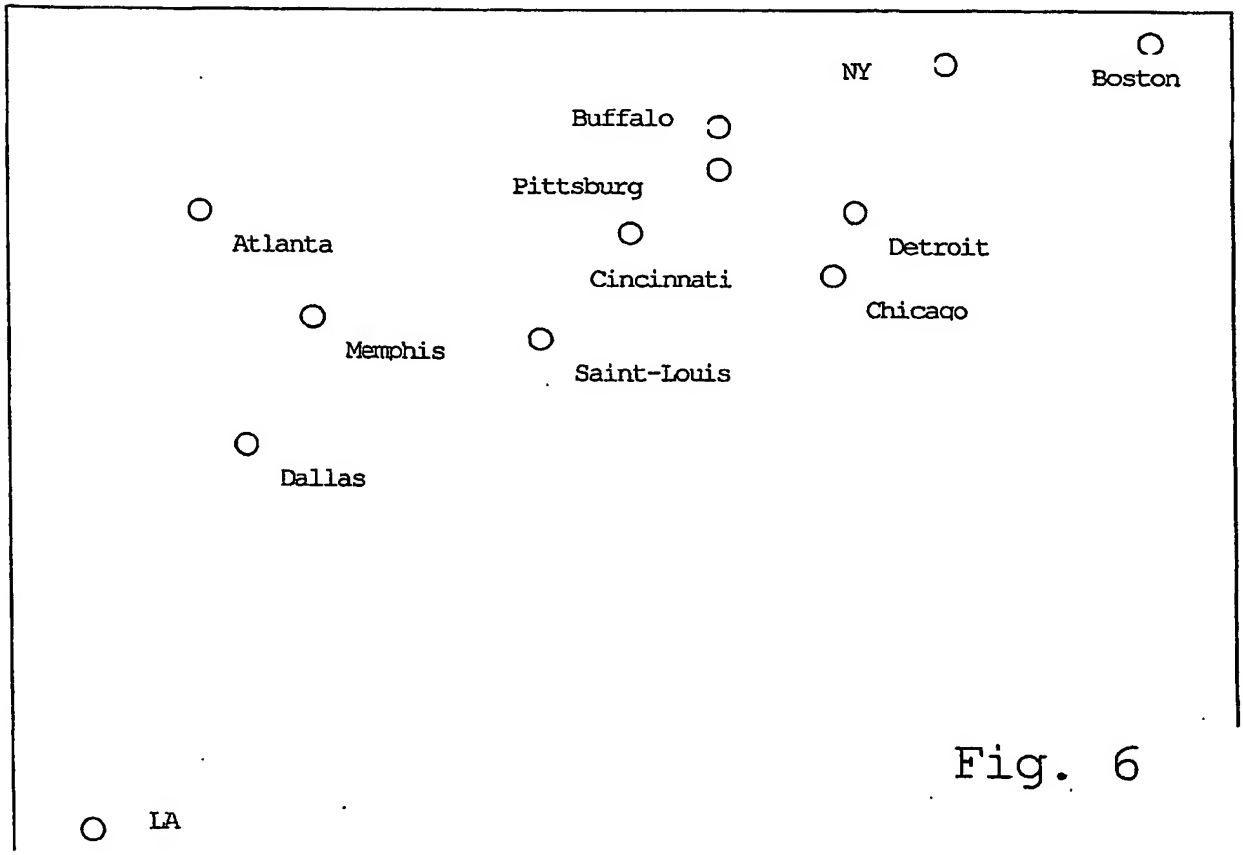


Fig. 6

Example 3

	Cereals	Rice	Potatoes	Sugar	Vegetables	Meat	Milk	Butter	Eggs
elgium	72,2	4,2	98,8	40,4	103,2	102	80	7,7	14,2
enmark	70,5	2,2	57	39,5	50	105,8	145,2	4,1	14,3
ermany	71,3	2,3	74,1	37,1	83,1	97,2	90,7	6,9	14,8
reece	109,8	5,4	90	30	229,5	77,1	63,1	0,9	11,3
Spain	71,4	5,8	107,8	26,8	191,7	102,1	98,4	0,6	15,3
rance	73	4,3	78,2	34,1	95	110,5	98,9	8,9	15
eland	93,4	3,2	151,5	34,8	55	105	185,9	3,4	11,4
Italy	110,2	4,8	38,6	27,9	181,9	88	65	2,4	11,1
herland	54,6	5	86,7	39,7	99	89,4	136,2	5,4	10,7
ortugal	86	5,7	106,6	29,4	100	75,5	96	1,5	7,7
Britain	74,3	4,5	94,1	39,8	60	74,4	129,3	3,2	10,8
ustria	68,7	4,2	62,6	37,1	81,9	93,4	121,3	4,3	13,4
nland	70,1	5,4	61,6	35,7	52,6	65	208,4	5,8	10,9
sland	79,7	1,9	50,2	54,9	50	71,7	205,6	4,6	11,3
orway	76,9	3,5	73,2	37,3	48,3	54,9	176,5	2,1	11,3
weden	69,3	4,3	70	37,5	48,5	60,5	154,1	5,7	12,9

European
Countries
Food
Consumption in
1994:

9 variables
16
observations

Fig. 7

Example 3

	Cereals	Rice	Potatoes	Sugar	Vegetables	Meat	Milk	Butter	Eggs
Belgium	72,2	4,2	98,8	40,4	103,2	102	80	7,7	14,2
Denmark	70,5	2,2	57	39,5	50	105,8	145,2	4,1	14,3
Germany	71,3	2,3	74,1	37,1	83,1	97,2	90,7	6,9	14,8
Greece	109,8	5,4	90	30	229,5	77,1	63,1	0,9	11,3
Spain	71,4	5,8	107,8	26,8	191,7	102,1	98,4	0,6	15,3
France	73	4,3	78,2	34,1	95	110,5	98,9	8,9	15
Ireland	93,4	3,2	151,5	34,8	55	105	185,9	3,4	11,4
Italy	110,2	4,8	38,6	27,9	181,9	88	65	2,4	11,1
Netherland	54,6	5	86,7	39,7	99	89,4	136,2	5,4	10,7
Portugal	86	5,7	106,6	29,4	100	75,5	96	1,5	7,7
Gr Britain	74,3	4,5	94,1	39,8	60	74,4	129,3	3,2	10,8
Austria	68,7	4,2	62,6	37,1	81,9	93,4	121,3	4,3	13,4
Finnland	70,1	5,4	61,6	35,7	52,6	65	208,4	5,8	10,9
Island	79,7	1,9	50,2	54,9	50	71,7	205,6	4,6	11,3
Norway	76,9	3,5	73,2	37,3	48,3	54,9	176,5	2,1	11,3
Sweden	69,3	4,3	70	37,5	48,5	60,5	154,1	5,7	12,9

European Countries Food Consumption in 1994:

9 variables
16 observations

Fig. 7

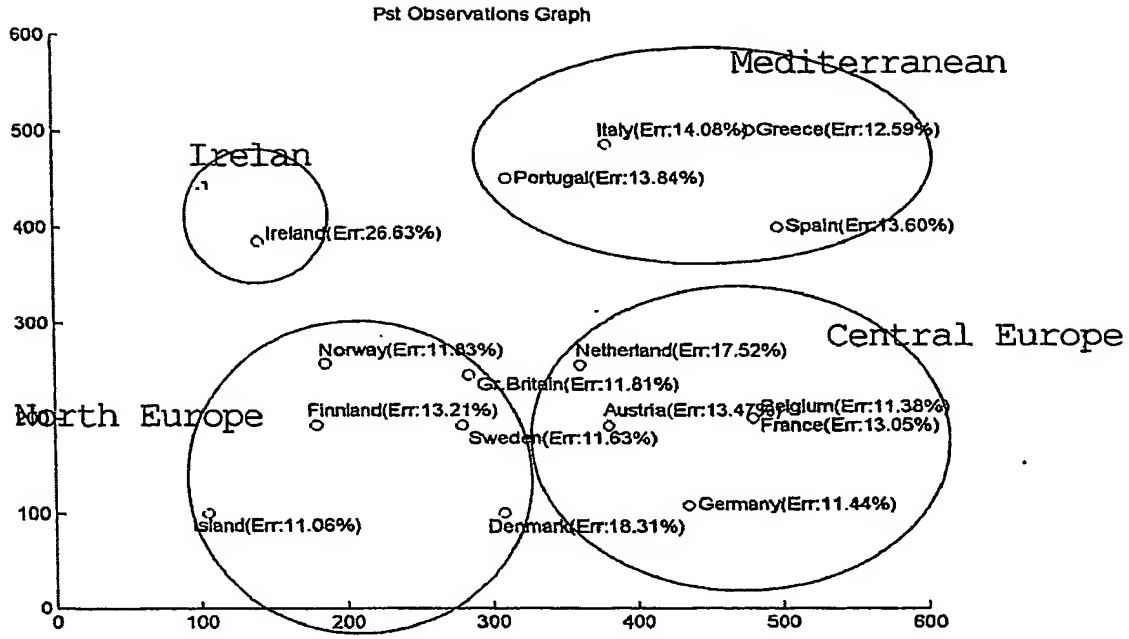

Pst Observations Graph
Mediterranean
Irelan
Central Europe
North Europe
Italy(Err:14.08%)
Greece(Err:12.59%)
Portugal(Err:13.84%)
Spain(Err:13.60%)
Ireland(Err:26.63%)
Norway(Err:11.83%)
Netherland(Err:17.52%)
Gr.Britain(Err:11.81%)
Finnland(Err:13.21%)
Austria(Err:13.47%)
Belgium(Err:11.38%)
France(Err:13.05%)
Sweden(Err:11.63%)
Germany(Err:11.44%)
Island(Err:11.06%)
Denmark(Err:18.31%)
600
500
400
300
100
0
0
100
200
300
400
500
600

Fig. 8

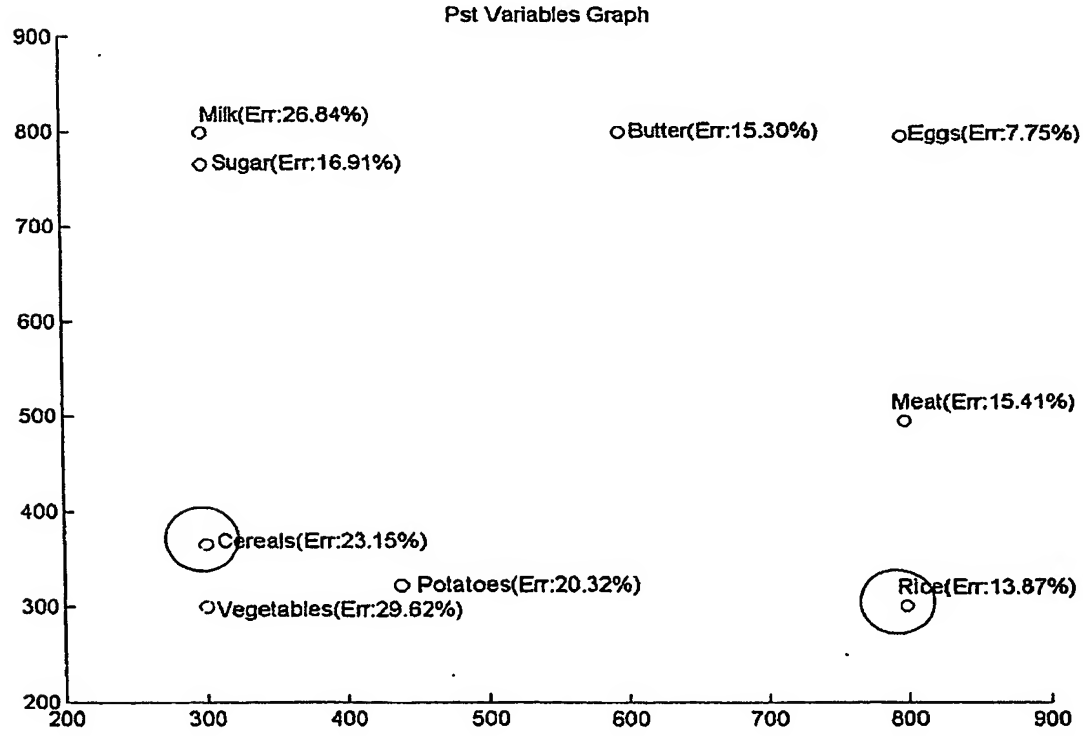

Pst Variables Graph
Milk(Err:26.84%)
Butter(Err:15.30%)
Eggs(Err:7.75%)
Sugar(Err:16.91%)
Meat(Err:15.41%)
Cereals(Err:23.15%)
Potatoes(Err:20.32%)
Rice(Err:13.87%)
Vegetables(Err:29.62%)
900
800
700
600
500
400
300
200
200
300
400
500
600
700
800
900

Fig. 9

Variables		Complement
1	AgeExam	1-AgeExam
2	AgeDeath	1-AgeDeath
3	EdYears	1-EdYears
4	ADL	1-ADL
5	WRCL	1-WRCL
6	CNPR	1-CNPR
7	BOST	1-BOST
8	VRBF	1-VRBF
9	MMSE	1-MMSE
10	TangleNeocortex	1-TangleNeocortex
11	TangleHippo	1-TangleHippo
12	PlaqueNeocortex	1-PlaqueNeocortex
13	PlaqueHippo	1-PlaqueHippo

Fig. 10

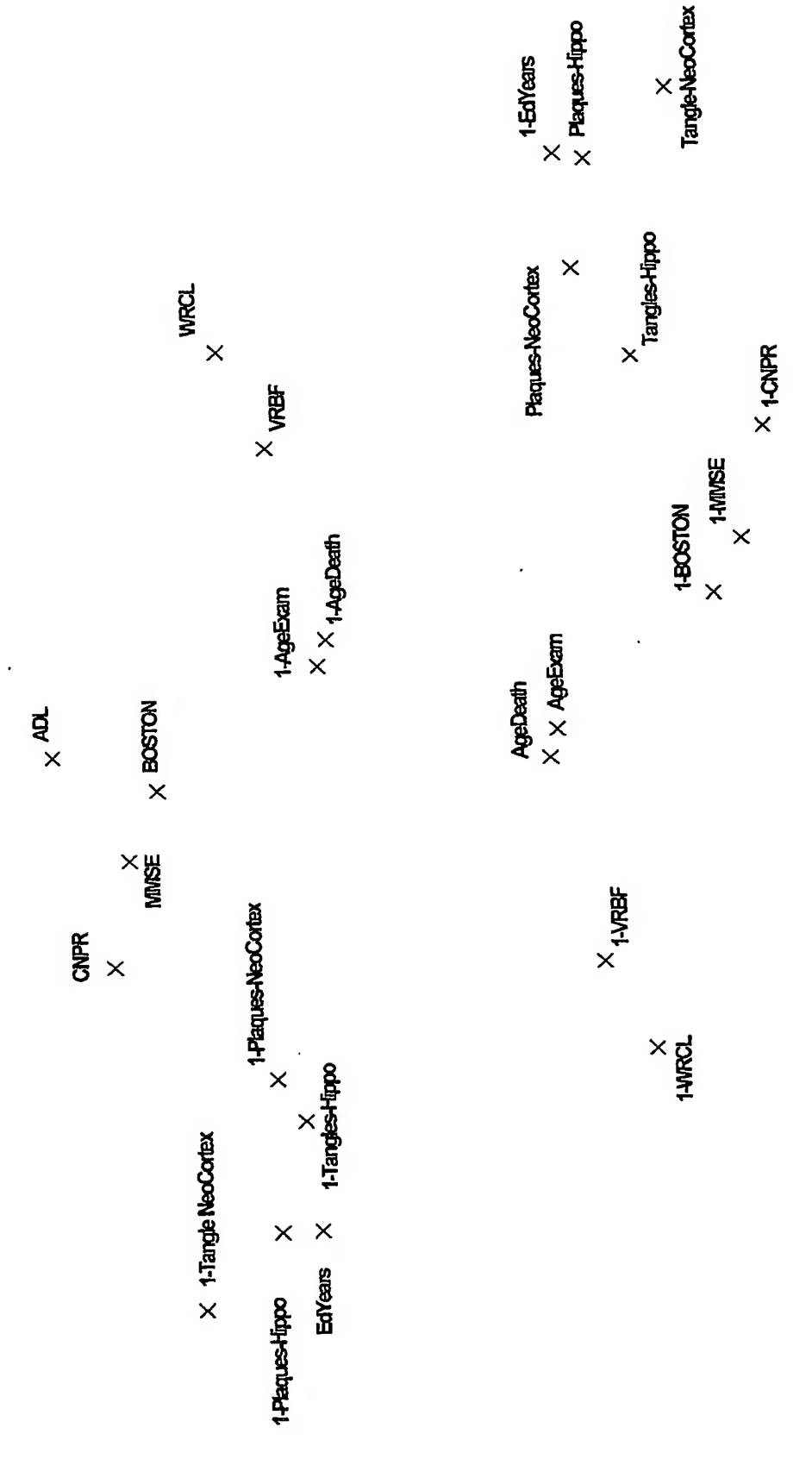

ADL
BOSTON
MMSE
CNPR
WRCL
VRBF
1-AgeExam
1-AgeDeath
1-Plaques-NeoCortex
1-Tangles-Hippo
1-Tangle NeoCortex
1-Plaques-Hippo
EdYears
AgeDeath
AgeExam
1-VRBF
1-WRCL
1-EdYears
Plaques-Hippo
Plaques-NeoCortex
Tangles-Hippo
Tangle-NeoCortex
1-CNPR
1-BOSTON
1-MMSE
1-ADL

Fig.11

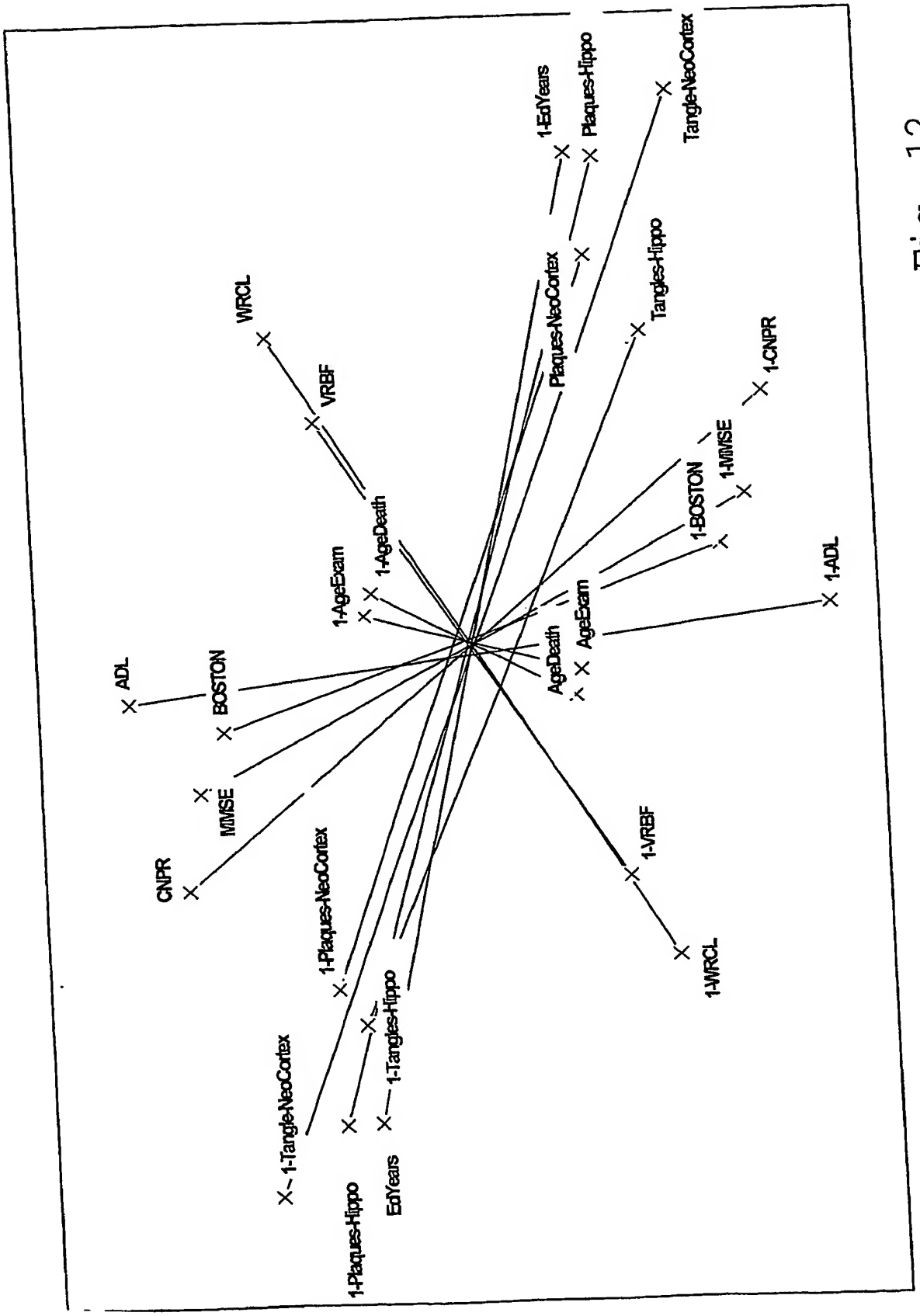

WRCL
VRBF
1-AgeExam
1-AgeDeath
ADL
BOSTON
MMSE
CNPR
1-Plaques-NeoCortex
1-Tangles-Hippo
1-Tangle-NeoCortex
1-Plaques-Hippo
EdYears
1-EdYears
Plaques-Hippo
Plaques-NeoCortex
Tangles-Hippo
Tangle-NeoCortex
AgeDeath
AgeExam
1-BOSTON
1-MMSE
1-CNPR
1-ADL
1-VRBF
1-WRCL

Fig. 12

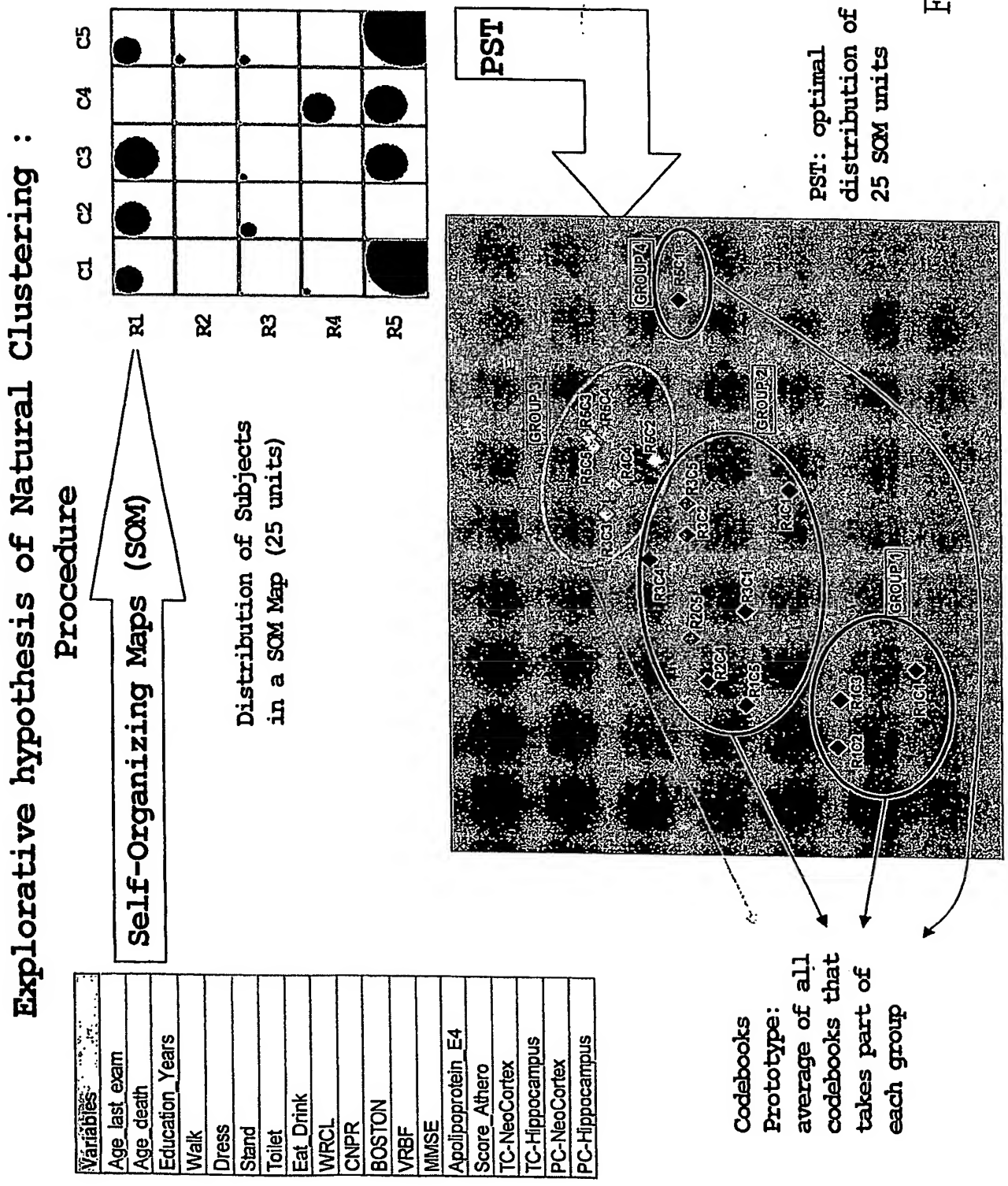
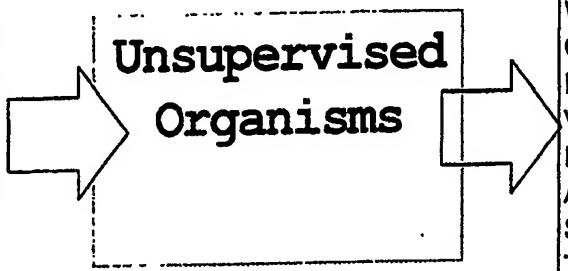

Explorative hypothesis of Natural Clustering :
Procedure
Variables
Age_last_exam
Age_death
Education_Years
Walk
Dress
Stand
Toilet
Eat_Drink
WRCL
CNPR
BOSTON
VRBF
MMSE
Apolipoprotein_E4
Score_Athero
TC-NeoCortex
TC-Hippocampus
PC-NeoCortex
PC-Hippocampus
Self-Organizing Maps (SOM)
Distribution of Subjects
in a SOM Map (25 units)
C1 C2 C3 C4 C5
R1 R2 R3 R4 R5
PST
PST: optimal
distribution of the
25 SOM units
GROUP 1
GROUP 2
GROUP 3
GROUP 4
Codebooks
Prototype:
average of all
codebooks that
takes part of
each group

Fig. 13

Explorative hypothesis of Natural Clustering

Variables
Age_last_exam
Age_death
Education_Years
Walk
Dress
Stand
Toilet
Eat_Drink
WRCL
CNPR
BOSTON
VRBF
MMSE
Apolipoprotein_E4
Score_Athero
TC-NeoCortex
TC-Hippocampus
PC-NeoCortex
PC-Hippocampus



Codebooks Prototype of each Group

Variables	Group-1	Group-2	Group-3	Group-4
Age_last_exam	90,7579	89,3621	84,9923	88,1756
Age_death	91,4622	90,1110	86,0971	89,1812
Education_Years	14,5917	14,2293	15,2572	16,0271
Walk	0,0605	0,5899	0,9713	0,9768
Dress	0,0453	0,4409	0,9791	0,9466
Stand	0,0652	0,5912	0,9913	1,0000
Toilet	0,0041	0,1725	0,9560	0,9780
Eat_Drink	0,0280	0,6173	0,8813	0,9347
WRCL	0,2860	2,2480	5,3737	3,4404
CNPR	2,5170	8,0884	9,5579	9,6409
BOSTON	3,1969	9,4987	11,8490	10,6115
VRBF	1,9532	7,3083	12,2307	12,5998
MMSE	4,7406	17,5601	25,7396	23,2561
Apolipoprotein_E4	0,3323	0,1622	0,0076	0,8658
Score_Athero	0,4845	0,4546	0,3899	0,5456
TC-NeoCortex	15,5941	7,7160	0,9555	7,2049
TC-Hippocampus	39,0581	33,3978	11,9865	31,9796
PC-NeoCortex	8,2940	5,7019	3,6005	6,3076
PC-Hippocampus	4,0608	3,0796	0,9535	5,1484
Number of subjects	**27**	**22**	**50**	**18**
Number of Demented	**25**	**14**	**6**	**1**
Demented in %	**92,59%**	**63,64%**	**12,00%**	**5,56%**
Number of MCI	**1**	**6**	**15**	**7**
MCI in %	**3,70%**	**27,27%**	**30,00%**	**38,89%**

Fig. 14